Foundations of Probability and Statistics
2020-2021
Analisi sull'indagine sui percorsi di studio e di lavoro dei diplomati

Daniele Rizzo (872359), Lorenzo Doneda (806727), Alessandro Maccario (865682)

Abstract

L'indagine sui percorsi di studio e di lavoro dei diplomati delle scuole secondarie di secondo grado è un'importante rilevazione svolta periodicamente dall'ISTAT. L'analisi presentata vuole rispondere ad alcune domande significative e ad alcuni preconcetti che si hanno, spesso, in merito alla provenienza sociale degli alunni, al loro rapporto con gli insegnanti e al reddito ottenuto a seconda della scelta universitaria. Ad esempio: quanto è importante provenire da un contesto familiare in cui i genitori possiedono un titolo di studi universitario rispetto ad un contesto in cui i parenti possiedono un titolo di studio elementare? I risultati dei voti di diploma fra genere maschile e femminile sono significativamente differenti? Si ottiene una significativa differenza in termini reddituali a seconda del tipo di titolo universitario ottenuto?

1 INTRODUZIONE

Questo lavoro presenta i dati relativi all'indagine "percorsi di studio e di lavoro dei diplomati' ' nel periodo di riferimento 2015. L'intento è quello di analizzare l'ambito scolastico e lavorativo degli intervistati considerando differenti variabili sia qualitative che quantitative.

I dati analizzati possono essere reperiti tramite il sito dell'Istituto Nazionale di Stastica al cui indirizzo è possibile ottenere il relativo dataset, la nota metodologica e la lista delle variabili.

I dati rilevati permettono di analizzare comparativamente la resa dei diversi titoli di studio sul mercato del lavoro rappresentando uno strumento per valutare l'efficacia complessiva dei sistemi di istruzioni superiore. Tali informazioni consentono inoltre di studiare quale relazione esista, se esiste, tra l'origine sociale e il processo di selezione scolastica e universitaria, oltreché il ruolo nel reddito ottenuto a livello lavorativo, ma anche le differenze di genere in termini di resa scolastica sui voti di diploma.

I dati relativi all'indagine completa sono stati rilevati dall'ISTAT nell'anno 2015 facendo riferimento alla leva scolastica del 2011. Per la somministrazione del questionario si è scelta la tecnica mista CAWI-CATI (Computer Assisted Web Interviewing e Computer Assisted Telephone Interviewing).

I contenuti informativi del questionario hanno riguardato tre aree tematiche di interesse:

- Gli studi
- La formazione professionale
- La situazione familiare

L'indagine effettuata è stata di tipo campionario a due stadi di selezione con stratificazione delle unità di primo stadio (primo stadio: le unità scolastiche; secondo stadio: gli alunni che hanno conseguito il diploma nell'anno solare 2010-2011). Sono presenti 26.235 osservazioni e 161 variabili.

Per lo studio da noi effettuato verranno però considerate solamente 10 di queste, ovvero:

- Sesso
- Cittadinanza
- Scuola pubblica/privata
- Voto del diploma
- Tipo di corso
- Area disciplinare
- Corso desiderato
- Reddito mensile totale netto
- Titolo di studio del padre
- Titolo di studio della madre

Dove si avranno variabili:

- Qualitative nominali: sesso (binaria), cittadinanza, scuola pubblica o privata (binaria), corso desiderato (binaria), tipo di corso, area disciplinare universitaria;
- Qualitative ordinali: titolo di studio del padre, titolo di studio della madre;
- Quantitative discrete: voto diploma;
- Quantitative continue: reddito mensile netto.

2 ANALISI ESPLORATIVA DEI DATI

A causa dell'elevato numero di valori mancanti presenti nelle variabili prese in considerazione, si è deciso di eliminare le osservazioni corrispondenti in quanto, per alcuni attributi, rappresentavano oltre il 50% del totale dell'attributo.

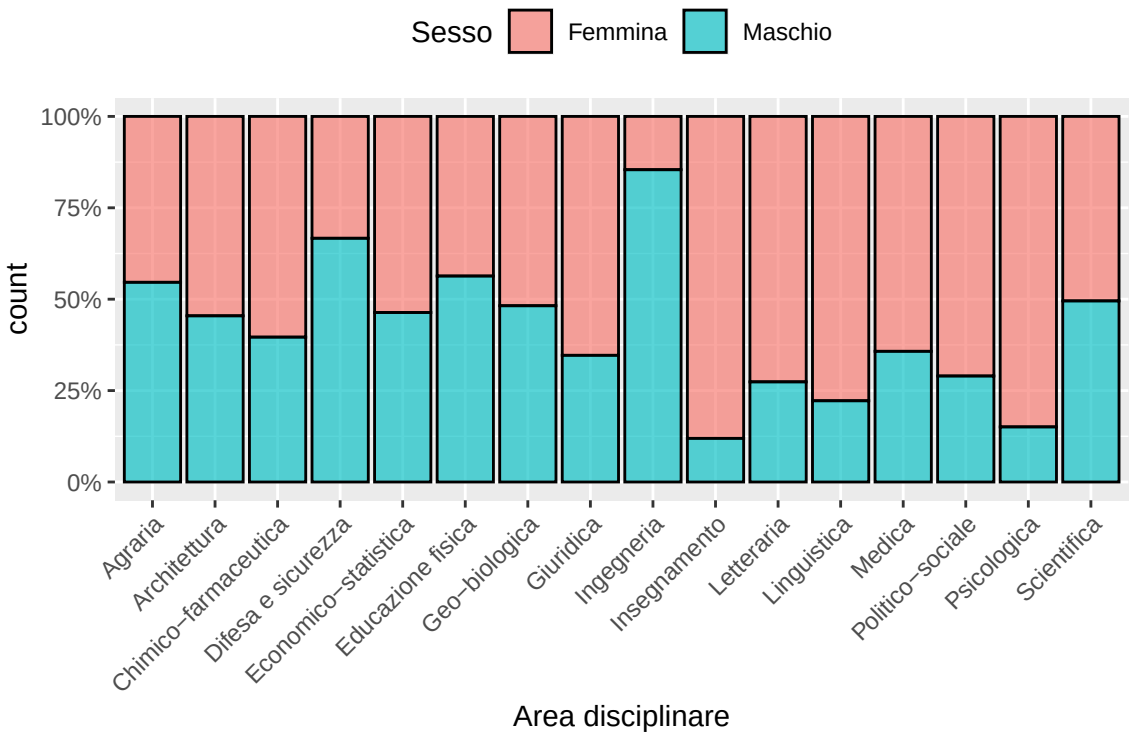
Per motivi di analisi inoltre, sono stati considerati solamente quegli studenti che hanno per scelta deciso di proseguire i loro studi a livello universitario.

Delle 26.235 osservazioni se ne analizzeranno quindi 3.089, considerate comunque sufficienti per gli scopi della seguente trattazione.

Prima ancora di addentrarsi nelle specifiche analisi, è utile poter riassumere quelle che sono le caratteristiche delle variabili prese in esame.

Poiché molte sono le variabili di tipo categoriale, l'output non potrà ritornare i valori minimi o massimi. Al contrario, invece, del caso in cui la variabile sia di tipo numerico, come per il *voto del diploma* o del *reddito mensile totale netto*. Si vuole sottolineare come siano presenti valori massimi fortemente distanti dal terzo quartile come nel caso del reddito mensile totale netto (2200), potendo indicare la presenza di possibili outliers maggiormente visibili tramite una rappresentazione a boxplot.

Una prima analisi che può essere svolta è relativa alla divisione per genere nelle varie aree disciplinari universitarie per visualizzare quale sia la distribuzione per disciplina universitaria:



Come è chiaramente visibile, si nota una forte percentuale di studenti maschili nel campo degli studi ingegneristici mentre, al contrario, in campo psicologico, letterario, linguistico e dell'insegnamento la componente femminile è prevalente, con oltre il 70% di studentesse.

Per avere poi un'idea più specifica del livello di preparazione degli studenti e delle studentesse che si iscrivono ad un ciclo di studi universitario, è conveniente analizzare quale fosse, in media, il voto ottenuto alla conclusione degli studi di scuola superiore, oltreché la deviazione standard:

Femmina	Maschio
77.84	75.54

Femmina	Maschio
0.198	0.1939

Le studentesse presentano una media più alta di circa 2,3 punti ma anche una deviazione standard di poco superiore. Per avere una certezza maggiore sull'effettiva differenza tra le medie, è possibile effettuare un test per assicurarsi che la differenza in media sia diversa dal valore nullo, dandoci maggiore confidenza sul fatto che esse siano effettivamente diverse (e che non lo siano per puro caso):

Table 3: t.test per la differenza delle medie dei voti

Test statistic	df	P value	Alternative hypothesis	mean in group Femmina	mean in group Maschio
5.736	2644	1.08e-08 *	two.sided	77.84	75.54

Il valore del p-value pari a 0 ci fornisce una buona confidenza nelle nostre ipotesi: le differenze fra le medie sono statisticamente significative.

Può essere inoltre analizzata la dipendenza tra il voto di diploma dei figli e la scolarizzazione dei genitori. Tale azione può essere intrapresa dapprincipio guardando all'eventuale dipendenza fra le due variabili, ma ci si soffermerà solo sul quella fra voto di diploma dei figli e istruzione del padre:

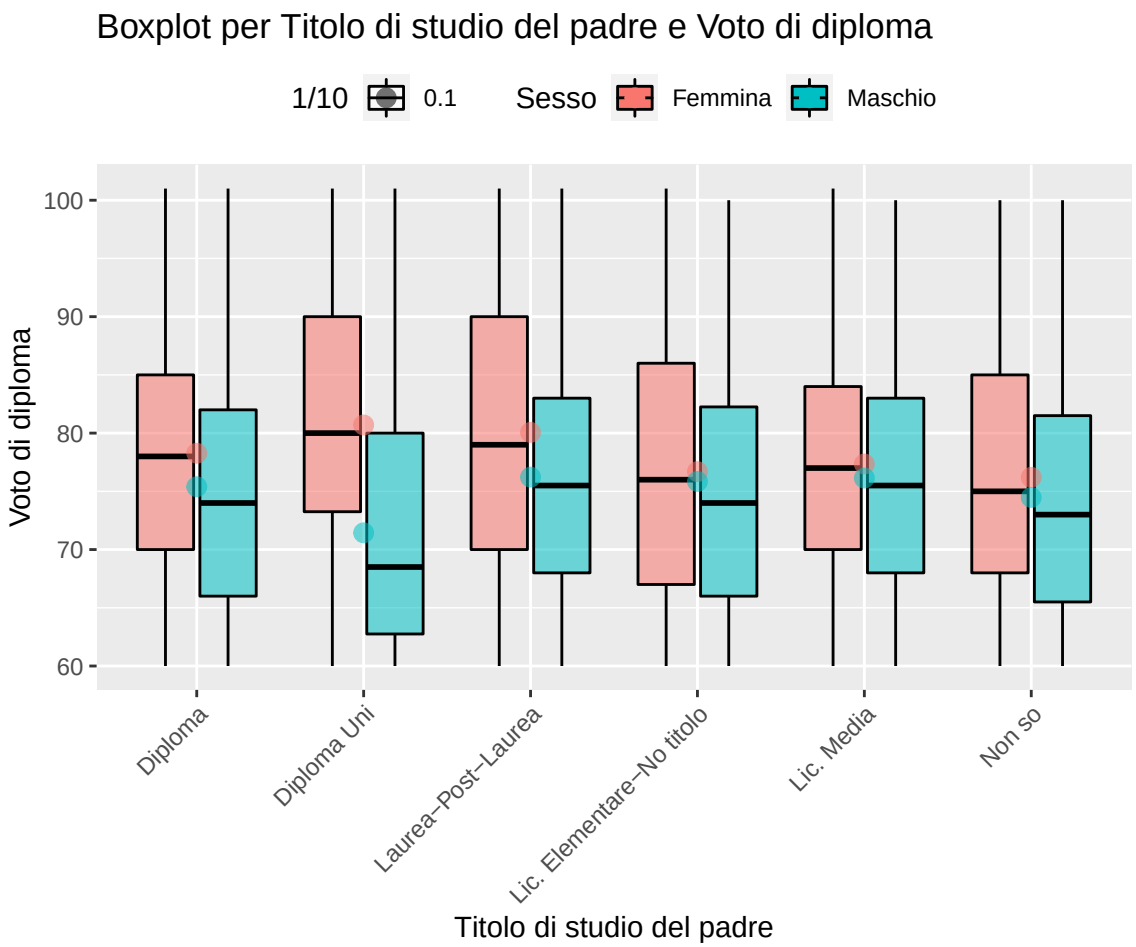
Number of cases in table: 3089 Number of factors: 2

Table 4: Test per l'indipendenza fra voto di diploma e titolo di studio del padre

Chisq	df	p.value
207	205	0.4469

Chi-squared approximation may be incorrect

Dai risultati del test, non sembra possibile rifiutare l'ipotesi nulla di indipendenza delle variabili. Possiamo rappresentare tramite boxplot tale relazione per essere maggiormente sicuri dei risultati analitici ottenuti:



Si può notare una forte differenza della media nel caso in cui il titolo di studio del genitore paterno sia il diploma di laurea: poiché però il test sulla dipendenza ha dato risultati negativi, è possibile che queste effettive differenze siano

dovute ad altri fattori che non sono stati presi in considerazione oppure, semplicemente, alla casualità del campione scelto.

La successiva domanda ha riguardato il voto di diploma in base all'aver frequentato o meno una scuola pubblica. Una scelta importante a livello genitoriale può essere proprio la scelta fra l'una o l'altra tipologia di istruzione. Una scuola privata fornisce una migliore istruzione che si potrebbe tradurre in una migliore possibilità di trovare lavoro con maggiore facilità e con un reddito più alto?

Privata	Pubblica
77.08	74.4

Privata	Pubblica
0.1979	0.1829

Il voto medio tende ad essere maggiore nelle scuole private con una deviazione standard però più alta rispetto a quella delle scuole pubbliche.

Si applica anche in questo caso un test sulle medie per esser più confidenti che il risultato precedentemente ottenuto non sia dovuto al caso:

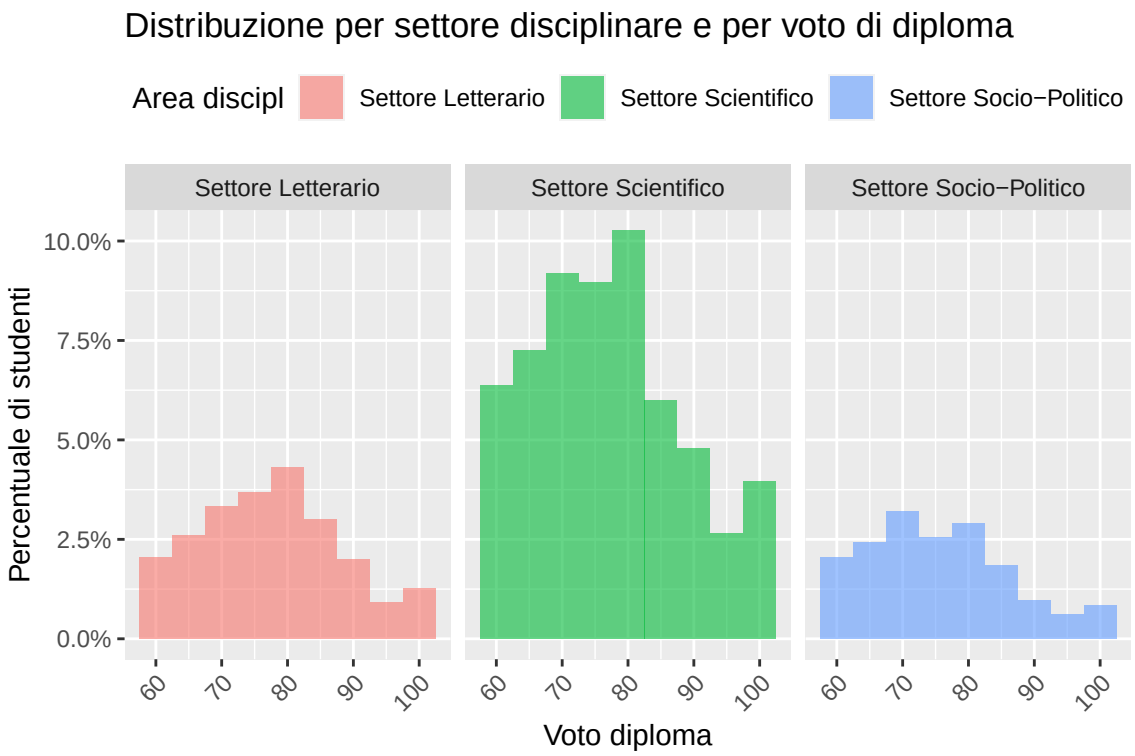
Table 7: t.test per la differenza fra le medie dei voti di diploma fra scuola pubblica e privata

Test statistic	df	P value	Alternative hypothesis	mean of x	mean of y
3.352	195.4	0.0009642 * * *	two.sided	77.08	74.4

Il valore del p-value uguale a 0 ci permette di rifiutare con buona certezza l'ipotesi nulla di differenza delle medie pari al valore nullo, quindi alla loro non uguaglianza: effettivamente la scuola privata tende a fornire migliori voti scolastici di conclusione delle scuole superiori.

Infine, può essere interessante anche capire come si distribuiscono gli studenti con voto più alto nelle diverse aree del sapere universitario. Chi ottiene risultati migliori alle scuole superiori, tende a scegliere un percorso di laurea diverso da chi, al contrario, non eccelle particolarmente? Le diverse facoltà sono state raggruppate in tre macro-aree principali nelle quali poi suddividere gli studenti appartenenti al dataset in analisi.

Settore Letterario	Settore Scientifico	Settore Socio-Politico
715	1836	538



Il maggior numero di studenti ha scelto un percorso universitario di tipo scientifico e fra questi il range di voti che viene più spesso rappresentato, corrispondente a circa il 10,2% del totale, è quello compreso fra circa 77 e 82. Gli studenti delle superiori i quali hanno ottenuto un voto maggiore di 95 sembrano aver scelto molto più frequentemente un settore scientifico di laurea.

3 REGRESSIONE LINEARE SUL VOTO DI DIPLOMA

Influenza dei genitori sul reddito dei figli e sul voto di diploma

E' un comune pensare che avere una certa istruzione in famiglia possa influire sia sull'istruzione dei figli e sui loro risultati accademici, ma anche sul reddito che questi possono ottenere. Si vuole quindi visulizzare l'eventuale dipendenza fra le variabili *Reddito mensile totale netto* e il *Titolo di studio del padre e della madre* per comprendere se effettivamente esista una dipendenza significativa, come anche fra il *voto di diploma* e l'istruzione dei genitori. Per farlo si applica un test del Chi-quadro:

- Dipendenza fra reddito mensile totale netto e titolo della madre: Number of cases in table: 3089 Number of factors: 2

Table 9: t.test per l'indipendenza fra reddito mensile totale netto e titolo della madre

Chisq	df	p.value
813.6	845	0.7755

Chi-squared approximation may be incorrect

- Dipendenza fra reddito mensile totale netto e titolo del padre: Number of cases in table: 3089 Number of factors: 2

Table 10: t.test per l'indipendenza fra reddito mensile totale netto e titolo del padre

Chisq	df	p.value
905.8	845	0.07198

Chi-squared approximation may be incorrect

- Dipendenza fra voto del diploma e titolo della madre: Number of cases in table: 3089 Number of factors: 2

Table 11: t.test per l'indipendenza fra voto di diploma e titolo della madre

Chisq	df	p.value
193.6	205	0.7048

Chi-squared approximation may be incorrect

- Dipendenza fra voto del diploma e titolo del padre: Number of cases in table: 3089 Number of factors: 2

Table 12: t.test per l'indipendenza fra voto di diploma e titolo del padre

Chisq	df	p.value
207	205	0.4469

Chi-squared approximation may be incorrect

Ciò che si può notare è che non si apprezzano particolari dipendenze, se non una lieve connessione tra il reddito mensile netto ed il livello d'istruzione del padre dell'intervistato a cui però non si vuole dare particolare attenzione (l'ipotesi di indipendenza può essere rifiutata solo con un livello di significatività dello $\alpha = 0.1\%$).

Tale risultato potrebbe però essere influenzato dal campione analizzato e non essere generalizzabile sull'intera popolazione. E' difatti presumibile che l'iscrizione all'unviversità degli studenti sia legata al grado di scolarizzazione genitoriale presente in famiglia.

Per quanto riguarda il reddito si osserva un fenomeno di indipendenza. In realtà bisogna considerare che la maggior propensione di iscrizione all'università potrebbe essere legata ai titoli di studio dei genitori e quindi, seguendo un percorso universitario, l'ingresso nel mondo del lavoro è posticipato e, di conseguenza, i redditi sono generalmente più bassi di chi invece ha iniziato a lavorare non appena finita la scuola secondaria superiore.

Reddito mensile totale netto e percorso universitario

Per valutare inoltre se, come viene comunemente pensato, l'ipotesi di seguire un percorso universitario non voluto sia necessario per ottenere alti redditi, sacrificandosi a corsi di laurea o percorsi di vita poco appassionanti sia corretta, si è deciso di analizzare il valore medio e mediano in termini di reddito, rispetto all'aver o meno seguito un percorso accademico di proprio interesse:

- Reddito mediano: *600*
- Reddito medio per chi ha frequentato un corso universitario non desiderato: *711*

I risultati ci dicono che in realtà non è proprio così. Lo stipendio mediano è pari a 600€ sia per chi ha seguito il corso desiderato che per chi non lo ha seguito. Gli studenti che hanno perseguito il percorso di studi voluto ha un reddito mensile netto di (717.87 $\pm$ 27.48)€ (C.L. 3σ), mentre per chi ha seguito un corso non desiderato di (704.34 $\pm$ 53.51)€ (C.L. 3σ). Questi risultati non permettono di affermare l'esistenza di una differenza tra queste medie che sia significativamente diversa da zero permettendo di di affermare che fare scelte non volute non necessariamente paga più di quanto non faccia quella che potrebbe essere il seguire una propria passione.

Rapporto con i professori e reddito mensile totale netto

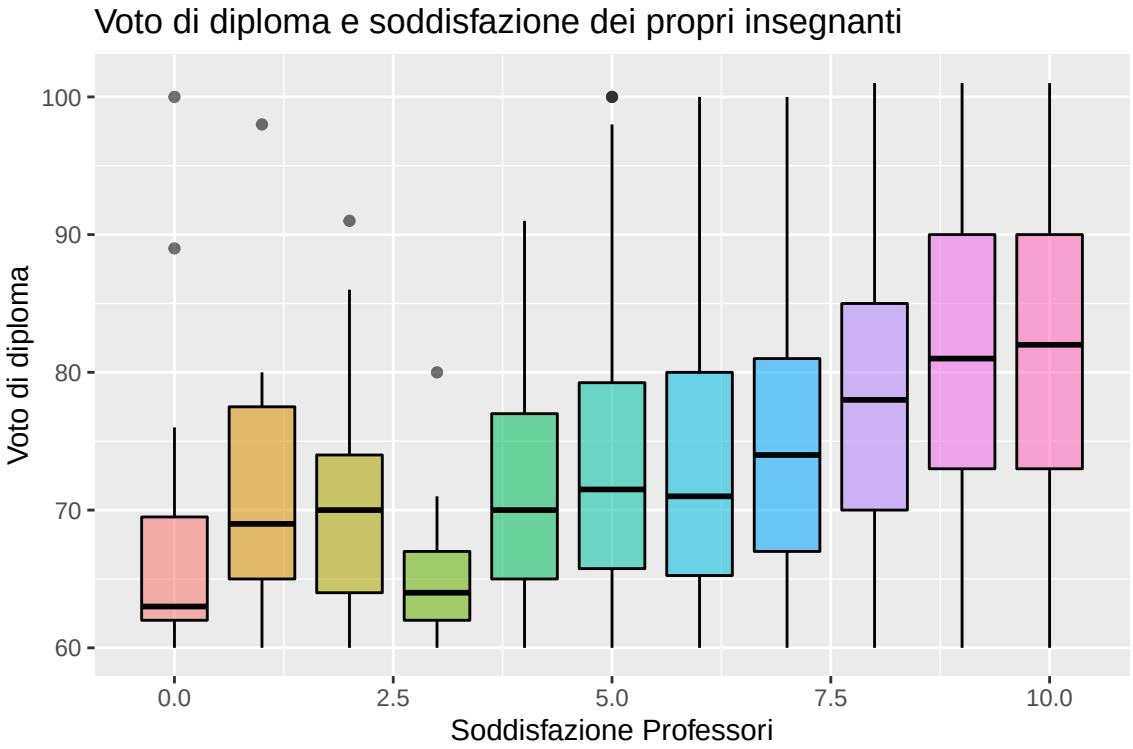
I dati a disposizione permettono inoltre di avere informazioni sulla valutazione del *rapporto con i professori*. E' di nostro interesse analizzare come il rapporto con i docenti possa influenzare sia i risultati nel breve periodo a livello scolastico, che sul lungo periodo a livello lavorativo. Viene utilizzato un test di indipendenza per valutarla tra i voti di diploma e il rapporto con gli insegnanti, oltreché una rappresentazione grafica:

Number of cases in table: 2973 Number of factors: 2

Table 13: t.test per l'indipendenza fra voto di diploma e soddisfazione del rapporto con gli insegnanti

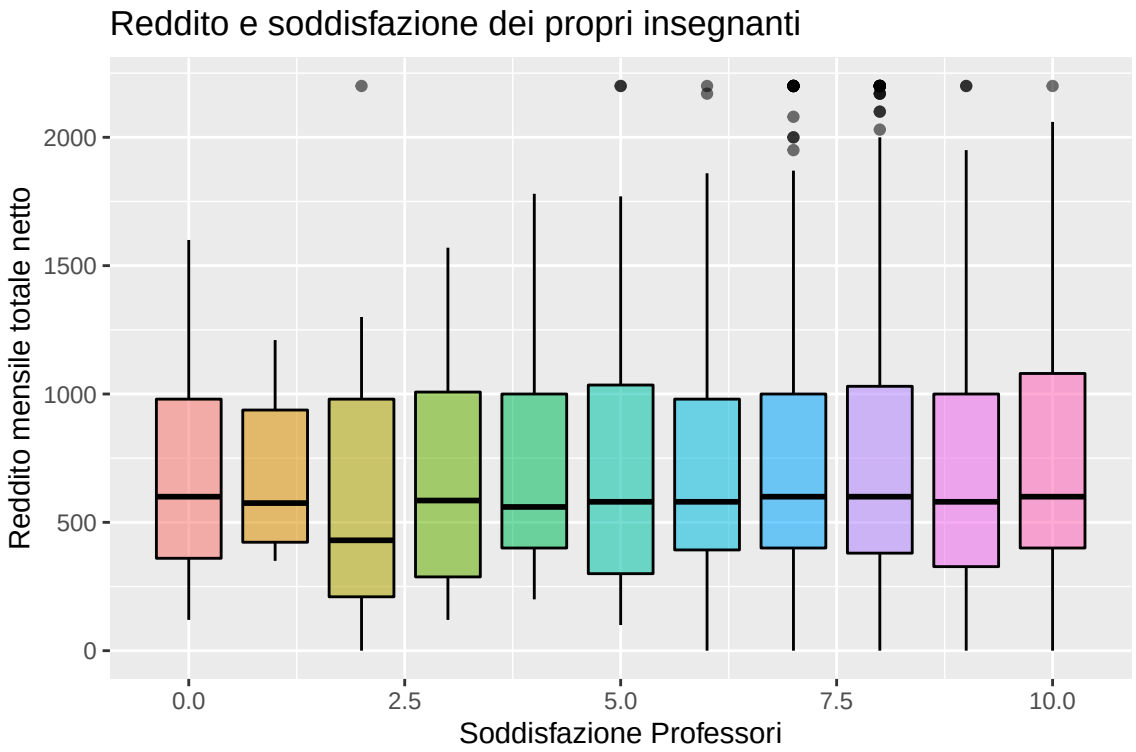
Chisq	df	p.value
654.2	410	1.713e-13

Chi-squared approximation may be incorrect



Valutando quindi i risultati scolastici, si nota come il rapporto con i professori sembra essere un elemento in forte connessione con il voto di diploma, visibile sia dal valore del p-value che permette di rifiutare l'ipotesi nulla di indipendenza, sia dal grafico a boxplot. In particolare, si può osservare nel boxplot come gli studenti che hanno incontrato una migliore esperienza con i professori siano stati anche quelli con dei voti di diploma generalmente più alti. Questo si potrebbe collegare al fatto che una maggior soddisfazione sia legata ad una miglior capacità dei docenti di comunicare efficacemente la propria conoscenza agli alunni e di trasmettere così metodo e nozioni tali da permettere una più effettiva espressione accademica degli studenti. Ciò sottolinea come la figura dell'insegnante sia fondamentale nella crescita e che i risultati di uno studente o studentessa siano strettamente legati anche all'incontro con un bravo mentore.

A livello lavorativo invece, si è valutata la dipendenza esistente fra la soddisfazione con gli insegnanti e il reddito mensile totale netto, tramite una rappresentazione a boxplot.



Il reddito mensile risulta anch'esso dipendente dalla soddisfazione dell'insegnamento. E' ipotizzabile che un migliore rapporto discente-insegnante sia foriero di una scelta più oculata a livello universitario in base alle proprie effettive passioni e, quindi, ad un miglior rendimento accademico traducibile in un reddito più elevato rispetto ad una situazione opposta.

L'altra valutazione in esame è relativa alla partecipazione alle lezioni universitarie e al loro contributo alla retribuzione con, presumibilmente, migliori risultati lavorativi.

Number of cases in table: 3089 Number of factors: 2

Table 14: t.test per l'indipendenza fra reddito mensile totale netto e frequenza delle lezioni seguite

Chisq	df	p.value
387.8	338	0.0319

Chi-squared approximation may be incorrect

L'ipotesi di indipendenza tra la frequenza delle lezioni ed il reddito mensile può essere rifiutata con una significatività $\alpha = 0.05$ dato un p-value del test χ^2 pari a 0.0319. Questo aspetto sottolinea come in corsi di istruzione universitaria la presenza e, probabilmente, l'immergersi nell'ambiente universitario possa contribuire ad una maggior vicinanza con esso, traducendosi in migliori risultati anche sul lungo periodo e in diversi aspetti dell'attività lavorativa dell'individuo.

Regressione lineare multivariata

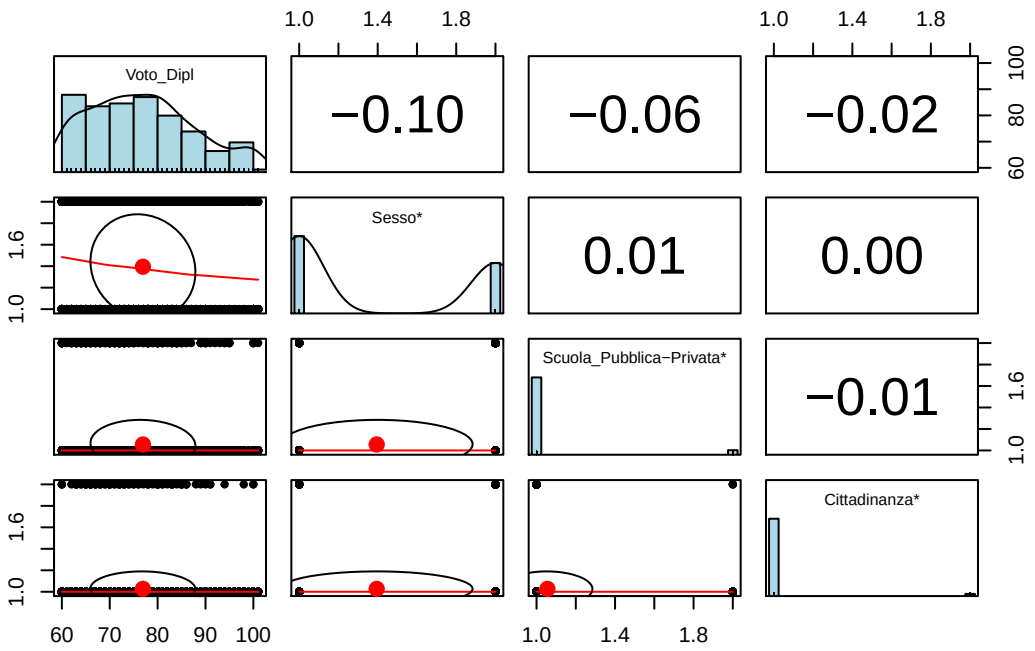
Dopo aver analizzato il dataset sotto differenti punti di vista, si è voluto successivamente provare a spiegare la variabile "Voto di diploma" in base ad altre variabili regressori come, ad esempio, l'aver o meno frequentato una scuola pubblica oppure il sesso dello studente (che, come già visto, sembra avere una certa influenza sul voto del diploma).

Per affrontare tale studio si è deciso di utilizzare una regressione lineare multivariata che può avere la duplice finalità di comprendere meglio le relazioni esistenti tra la variabile dipendente e le variabili indipendenti ma anche permettere di predire il valore del voto di diploma sulla base di determinati regressori scelti.

Una prima analisi a cui sottoporre le nostre variabili è quella relativa all'esistenza di multicollinearità delle variabili stesse: nel caso di elevata correlazione fra le variabili, i coefficienti di regressione possono risultare non corretti come anche le singole statistiche test 't'.

Questo può implicare che al variare, anche minimo, di una delle variabili esplicative, l'effetto sui coefficienti di regressione sia molto elevato; può accadere che il test F per la verifica di ipotesi calcolato sul complesso delle variabili sia significativo, ma che i singoli t test per i singoli parametri, al contrario, non lo siano.

Per procedere si guarda ai pairs.panels, rappresentazione che, sotto forma matriciale, indica le distribuzioni delle singole variabili, la curva di densità, e i singoli coefficienti di correlazione fra le variabili.



Le correlazioni lineari fra le variabili sono molto deboli permettendoci di escludere la presenza di multicollinearità.

Si decide quindi di utilizzare un modello di regressione lineare multivariata: come variabile dipendente si vuole usare il voto di diploma dello studente/studentessa e come regressori il sesso, l'aver frequentato una scuola pubblica o privata e la cittadinanza.

	Estimate	Std. Error	t value	Pr(>|t|)
Intercetta	78.03	0.2585	301.9	0
Sesso_Maschio	-2.283	0.4011	-5.692	1.37e-08
Scuola_Pubblica	-2.66	0.8553	-3.11	0.001888
Cittadinanza_Str-Apolide	-1.581	1.206	-1.311	0.1899

Table 16: Modello di regressione lineare applicato sulla variabile 'voto di diploma'

Observations	Residual Std. Error	R^2	Adjusted R^2
3089	10.9	0.01406	0.0131

I valori del p-value dei singoli regressori tende a confermare la significatività prevista dei regressori scelti, eccetto per la variabile *cittadinanza*: essa non è considerata significativa dal modello e quindi può essere esclusa dall'analisi. Il p-value della F-statistic allo stesso modo, conferma che tutti i regressori utilizzati sono considerabili statisticamente significativi. Si può inoltre notare quanto già visto nelle precedenti analisi: l'essere di sesso maschile implica avere un voto più basso di circa 2.28 volte rispetto ad essere di sesso femminile. Allo stesso modo, l'aver frequentato una scuola pubblica sembra affermare che abbia implicazioni negative sul voto di diploma.

Esaminando inoltre il valore dell'R^2, misura per valutare la bontà di adattamento del modello ai dati, si perviene ad un risultato scoraggiante se si vuole leggere il modello in termini predittivi: la variabilità che il modello riesce a spiegare è pari solo a circa l'$1,4$.

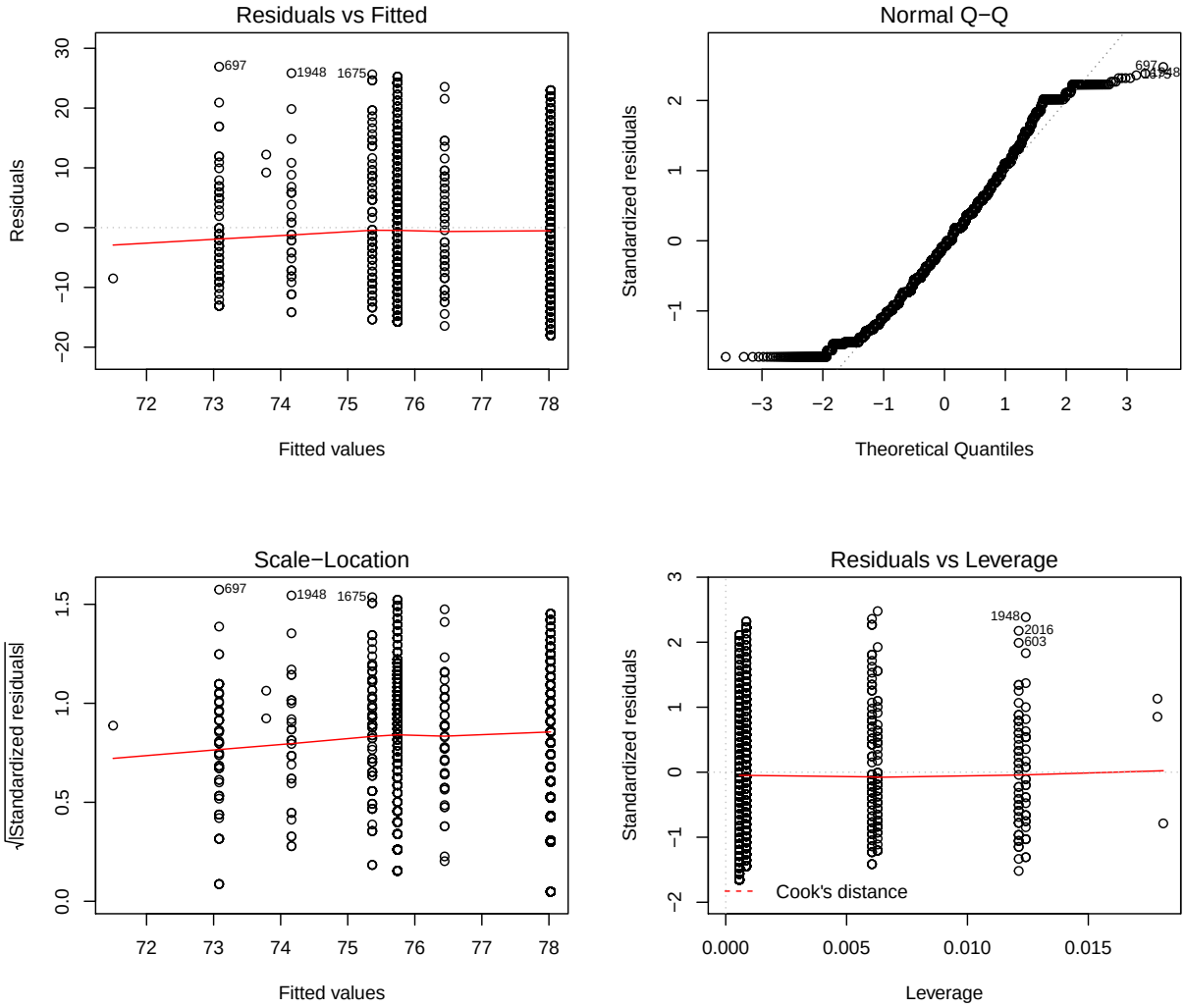
A questo punto, la domanda a cui si sarebbe dovuto rispondere fin dall'inizio sarebbe dovuta essere la seguente: i dati a disposizione permettono l'applicazione di tale modello? Ovvero, vengono soddisfatte le ipotesi alla base del modello di regressione lineare?

Poiché si tratta di un modello parametrico, si devono tenere in considerazione le ipotesi sulle quali esso si basa, ovvero:

- *Omoschedasticità degli errori*: gli errori devono avere la stessa varianza;
- *Indipendenza delle osservazioni*: esse sono state ottenute tramite metodi di campionamento statisticamente validi;
- *Normalità degli errori*: gli errori sono distribuiti normalmente;
- *Relazione tra la variabile dipendente e le indipendenti è di tipo lineare*: la linea di "best fit" attraversa i dati con una retta lineare (e non con una funzione curvilinea).

Per assicurarsi della validità di tali assunzioni si guarda al plot del modello di regressione lineare che permette di avere una visione d'insieme di quattro grafici fondamentali:

- *Residuals vs Fitted*: ci permette di visualizzare come si distribuiscono i residui intorno al valore nullo. Se fossero distribuiti randomicamente (ovvero, senza un preciso pattern) sarebbe possibile validare questo primo assunto;
- *Normal Q-Q plot*: tale visualizzazione ci permette di comprendere l'eventuale distribuzione normale dei residui rispetto ad una variabile casuale normale;
- *Standardized residuals vs Fitted values*;
- *Standardized residuals vs Leverage*.



Nel caso in esame, è visibile la non omoschedasticità degli errori che si dispongono con varianza crescente all'aumentare del voto di diploma. Il QQ-Plot inoltre, presenta delle code più pesanti dei residui rispetto alla distribuzione di una variabile casuale normale. Anche in questo caso, l'assunzione non è verificata. Poiché le prime due assunzioni non sono state incontrate, non si procede con la verifica degli altri grafici e si può affermare con più sicurezza che non sia presente una relazione di tipo lineare fra i dati in esame. Il modello di regressione lineare non potrà quindi essere considerato valido ai fini predittivi.

Per valutare analiticamente l'eventuale presenza o assenza di multicollinearità, si può infine sfruttare come altra misurazione il 'Fattore di inflazione della varianza' (*VIF - variance inflation factor*) per ciascuna variabile esplicativa.

Sesso	Scuola_pubblica-privata	Cittadinanza
1	1	1

L'interpretazione di tale indicatore è immediata in questo caso: poiché tutti i valori sono pressoché pari ad 1, si afferma che non sia presente multicollinearità nel modello scelto.

4 ANOVA ONE-WAY

Si vuole analizzare la variabile **Tipo di corso** in modo da suddividere la popolazione in cinque gruppi differenti: “Laurea Triennale”, “Laurea Magistrale”, “Laurea Ciclo Unico”, “Master” e “Università straniera” per verificare se le medie dei gruppi sono uguali fra loro.

Nelle due tabelle seguenti vengono riportate le frequenze assolute e relative in base al titolo di lurea conseguito.

Corso Uni Straniera	L. Ciclo Unico	L. Magistrale	L. Triennale	Master I Livello
25	348	327	2356	33

Corso Uni Straniera	L. Ciclo Unico	L. Magistrale	L. Triennale	Master I Livello
0.008093	0.1127	0.1059	0.7627	0.01068

Come si può evincere dai dati presentati precedentemente la maggioranza delle unità possiede una **Laurea triennale**, quasi il 77%, mentre la **Laurea magistrale** e la **Laurea a ciclo unico** sono state scelte da una percentuale compresa fra il 10% e il 12% degli studenti. Infine, le classi con la numerosità minore sono il **Master** e l’ **Università straniera**, con una percentuale attorno all’1%. La presenza elevata di valori nella classe **Laurea triennale** è condizionata dal presupposto che le rilevazioni sono state eseguite cinque anni dopo il conseguimento del diploma, il che comporta che i possessori dei titoli di studio con una durata di cinque anni, come **Laurea a ciclo unico** e **Laurea magistrale**, non hanno perso anni durante il loro percorso di studio universitario.

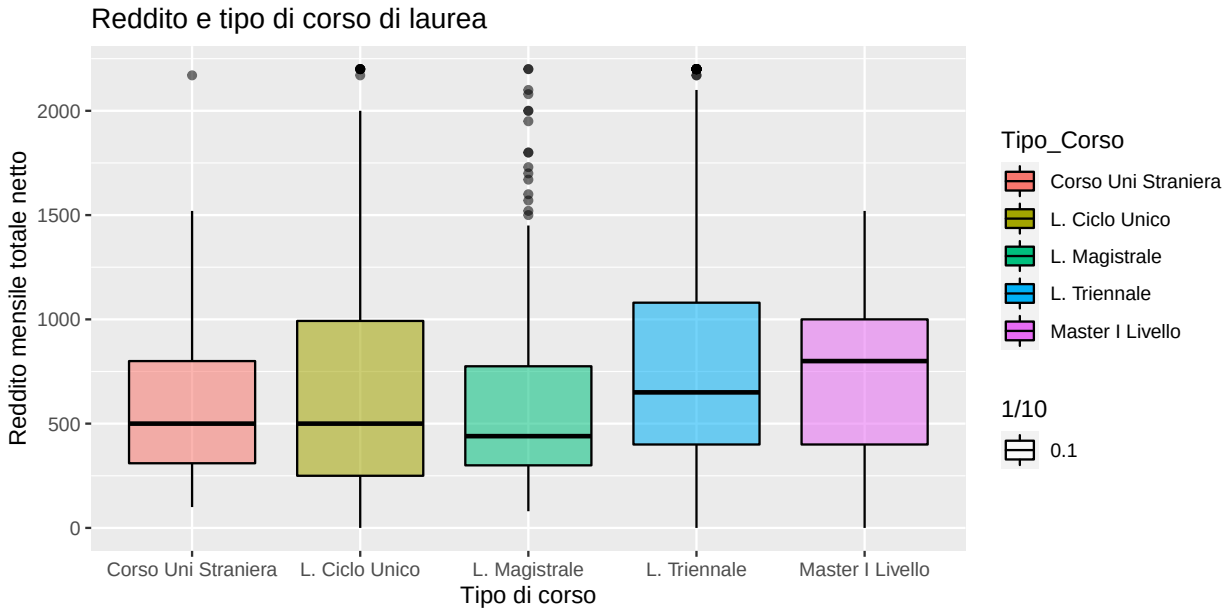
Nella seguente tabella verranno riportati i quartili, il minimo, il massimo, la media e la mediana degli stipendi a seconda del tipo di corso.

	Min.	First.Qu.	Median	Mean	Third.Qu.	Max.
Corso Uni Straniera	100	310	500	655	800	2,170
L. Ciclo Unico	0	250	500	652	992.5	2,200
L. Magistrale	80	300	440	573	775	2,200
L. Triennale	0	400	650	751	1,080	2,200
Master I Livello	0	400	800	722	1,000	1,520

Si può notare dalla media che i possessori di un titolo di **Laurea triennale** guadagnano di più, in media, rispetto agli altri gruppi: tale risultato è influenzato dalla diversa durata delle lauree in quanto la laurea triennale è un percorso che richiede meno tempo per il suo compimento rispetto agli altri e questa discrepanza si riflette direttamente sul reddito ottenuto dai laureati triennali.

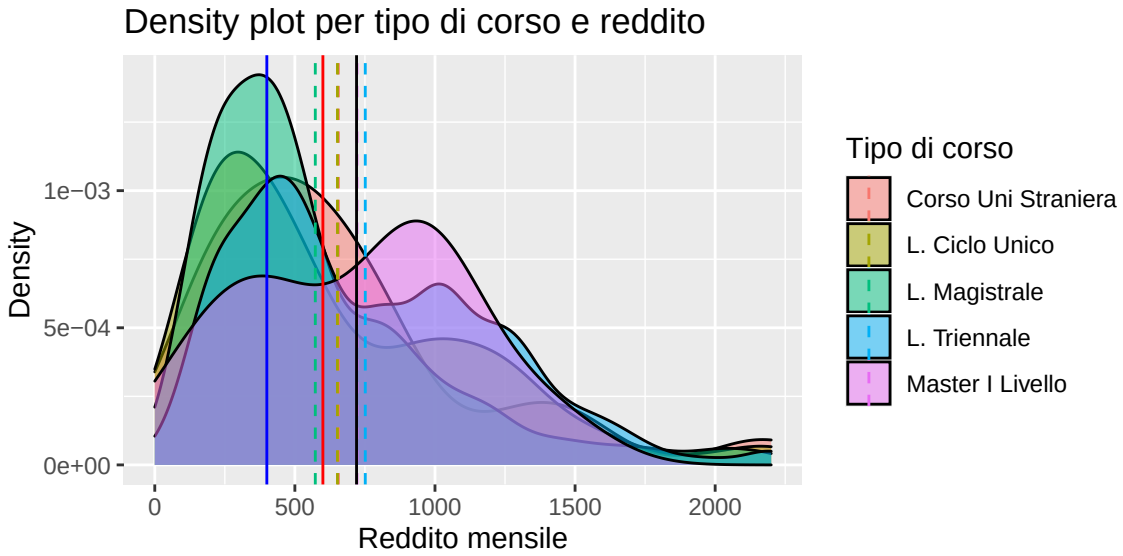
Se vengono osservati i valori riguardanti la mediana, il valore più alto si trova all’interno del gruppo del **Master** il che può far pensare alla presenza di outliers negli altri gruppi che influenzano direttamente il valore della media. Per i valori massimi non si notano differenze sostanziali ad esclusione per chi possiede un **Master**. Sui valori minimi si può notare la presenza di numeri molto bassi o pari a zero a causa, probabilmente, di contratti di stage spesso non retribuiti o pagati tramite un rimborso spese.

Per aiutare la comprensione dei dati precedentemente esposti viene anche inserita una rappresentazione grafica a box-plot in modo da verificare anche l’effettiva presenza di outliers.



Da questa analisi grafica si possono riprendere le affermazioni precedenti rispetto alla presenza degli outliers: difatti, nella classe **Master** non ve ne sono, non influenzando quindi la media. Com'era prevedibile non sono presenti outliers inferiori, mentre per gli outliers massimi, essi sussistono fortemente nelle classi di **Laurea Magistrale** e **Laurea triennale** a differenza degli altri titoli di studio. Il salario minimo delle diverse classi è maggiore per **Laurea magistrale** e **Università straniera**, mentre per le rimanenti classi è uguale a zero, ribadendo quanto specificato in precedenza sulla presenza di lavoro non retribuito.

Si visualizza anche un Density plot per comprendere meglio le differenze dello stipendio a seconda del titolo di studio.



Le linee rette intere blu, rossa e nera rappresentano rispettivamente la moda, la mediana e la media del reddito mensile nel complesso, senza considerare la suddivisione per titoli di studio. Le linee tratteggiate rappresentano, invece, la media di ogni classe. Dalla posizione della moda, della mediana e della media si può evincere che la curva di probabilità del **Reddito mensile totale netto** presenta un'asimmetria positiva con code sempre più leggere all'aumentare del

reddito: è ragionevole pensare che siano pochi gli individui che, concluso il ciclo universitario triennale o di master di primo livello, siano in grado di ottenere un reddito al di sopra dei 2000 euro.

Infine, è stata svolta un'analisi ANOVA sui sui diversi tipo di corso effettuando l'analisi in relazione allo stipendio, volendo dare risposta al quesito se, in media, il titolo di studio universitario possa portare effettivamente differenze nel reddito ottenuto:

Table 21: Test ANOVA One-Way

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Tipo_Corso	4	11,096,840	2,774,210	13.71	4.355e-11
Residuals	3,084	623,849,406	202,286	NA	NA

Il p-value rappresenta la probabilità di ottenere un valore maggiore rispetto al valore F-value, sapendo vera l'ipotesi nulla. Dato il valore sostanzialmente pari a 0, si può ragionevolmente affermare che non si presenta un'evidenza tale da far pensare che le medie dei gruppi relativi alla variabile **Tipo di corso** siano fra di loro uguali. Effettivamente, e ragionevolmente, si conclude che l'aver ottenuto un titolo di istruzione universitaria apporta benefici in termini reddituali significativamente diversi a seconda del tipo di corso.

CONCLUSIONI

Lo studio svolto ha visto quindi una prima esplorazione dei dati, al fine di comprendere la struttura del dataset sia a livello numerico che grafico. Si è cercato quindi di estrapolare eventuali relazioni fra il voto di diploma e l'accesso all'università, fra il sesso dello studente e il voto del diploma; fra l'aver conseguito il diploma in una scuola privata o pubblica e il voto finale.

L'analisi si è poi successivamente concentrata sull'applicazione del metodo di stima della regressione lineare per prevedere quale potesse essere il reddito netto ottenuto dai diplomati considerando differenti variabili esplicative.

Infine, si è applicato il test ANOVA one-way per valutare l'eventuale differenza, in media, tra il reddito ottenuto e il tipo di corso universitario scelto dagli studenti e l'esistenza o meno di una differenza in termini di reddito.

Si può affermare che esista una significativa differenza di voti scolastici al termine del ciclo di studi superiori fra il genere maschile e femminile dove, quest'ultimo, performa in media meglio dei colleghi maschi. Una percentuale consistente di circa il 10% di coloro che ottengono una votazione fra il 77 e l'82 decide di iscriversi ad una facoltà scientifica e fra coloro che ottengono votazioni ancora maggiori pari a circa 100 o 101, la maggioranza sceglie tali facoltà.

Dalle analisi svolte, si è notato che avere un buon rapporto con i propri insegnanti, tende a migliorare sia i voti scolastici degli alunni che, in parte, i redditi ottenuti a livello lavorativo per una, presumibile, maggiore capacità di eccellere negli studi universitari e di primeggiare fra i candidati assunti in azienda, con uno stupendio, fin da subito, più elevato.

La mancanza di un legame di tipo lineare fra la variabile voto di diploma e le variabili sesso, scuola pubblica-privata e cittadinanza impedisce di ottenere un modello attendibile di previsione del voto di diploma che sia affidabile in quanto non vengono incontrate le ipotesi che vi sono alla base.

Infine, può essere affermato che l'ottenimento della laurea triennale permetta l'ottenimento di un reddito più alto (almeno sul breve che è stato considerato) delle altre categorie di lauree considerate in quanto, per definizione, permette l'ingresso nel mondo del lavoro in anticipo accumulando un'esperienza lavorativa più ampia rispetto ai propri colleghi che decidono di proseguire con gli studi.

REFERENCES

ISTAT. 2015. "L'indagine Sui Percorsi Di Studio E Di Lavoro Dei Diplomati: Microdati Ad Uso Pubblico." 2015. https://www.istat.it/it/archivio/96042.

Moksony, Ferenc. n.d. "Small Is Beautiful. The Use and Interpretation of R2 in Social Research." https://www.academia.edu/3880005/Small_is_beautiful_The_use_and_interpretation_of_R2_in_social_research.

STHDA. n.d. "Quick Start Guide - R Software and Data Visualization." http://www.sthda.com/english/wiki/ggplot2-box-plot-quick-start-guide-r-software-and-data-visualization.